ENVIRONMETRICS
Environmetrics 2006; **17**: 363–385
Published online 19 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.775

Approximately optimal spatial design approaches for environmental health data

Gangqiang Xia[1], Marie Lynn Miranda[2] and Alan E. Gelfand[1]*,†

[1]*Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251, USA*
[2]*Nicholas School of the Environment and Earth Sciences, Duke University, Durham, NC 27708-0328, USA*

SUMMARY

Environmental health research considers the relationship between exposures to environmental contaminants and particular health endpoints. Due to spatial structure associated with either exposures or outcomes, spatial modeling is making rapid inroads in environmental health. Our focus in this paper is on approximately optimal spatial design in the case of one-time sampling at a large number of spatial locations. If we plan to use spatial processes in building models to analyze the data, it seems equally appropriate to use such models in developing the sampling design.

For a given study region, our contribution is to develop an approximately optimal sampling strategy to learn about the spatial distribution of a contaminant across the region. Optimal design, working with a continuum of locations, is intractable so, as is customary, we presume that the region has been gridded to high resolution. The criteria we focus on, are developed from the Fisher information matrix with the goal of learning not only about the regression structure in the model but also about the dependence structure.

Under a criterion that attempts to maximize information gain, we consider three strategies to develop an approximately optimal design—sequential sampling, block sampling, and stochastic search. We also discuss utility-based modification of these strategies to achieve oversampling with regard to specified objectives. We present some theoretical and empirical properties and relationships among these strategies and provide an illustrative implementation for a simulated dataset. We also describe a real application in the context of the toxics release inventory (TRI). Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: block sampling; entropy; Fisher information; Gaussian process; sequential sampling; stochastic search

1. INTRODUCTION

Environmental health research considers the relationship between exposure to environmental contaminants and particular health endpoints. Many environmental health issues are characterized by spatial structure in either the contaminant surfaces or the pattern of observed cases. Thus, spatial modeling is making rapid inroads in environmental health. For exposure, which is our focus, models

*Correspondence to: A. E. Gelfand, JB Duke Professor of Statistics and Decision Sciences ISDS, Duke University, Durham, NC 27708-0251, USA.
†E-mail: alan@stat.duke.edu

Received 26 June 2005
Revised 4 October 2005

that explicitly include spatial structure provide better explanation of contaminant surfaces both with regard to estimation of levels and the uncertainty in this estimation.

By definition, if exposure surfaces are envisioned as conceptually measurable at every (point) location in a study region, then such surfaces are inherently spatial in nature. Anticipating spatial association in contaminant levels, with an uncountable collection of locations, we naturally turn to point-referenced association models, that is, spatial process models. In this paper, our attention is to a particular aspect of sampling design: how shall we choose locations to sample exposure levels (possibly ambient or deposition) that are anticipated to be essentially static? For example, how shall we sample individuals within a region to measure contaminant levels in the blood? Or, how shall we sample locations to learn about ambient levels of air toxics or perhaps arsenic levels in the water table? We are focusing on one-time sampling at a large number of locations rather than designing long-term typically sparse monitoring networks. Thus, we are not considering the costs for installing, operating, and maintaining a network but rather the cost of collecting a single observation. If we plan to use spatial processes in building models to analyze such data, it seems equally appropriate to use such models in developing the sampling design.

Many GIS-based projects have been successful in examining the spatial nature of environmental health research and policy practice addressing toxic exposure, vector borne disease, health information access, and the built environment. See, for example, Miranda *et al*. (2002), Dolinoy and Miranda (2004), and further references therein. However, data collection exercises typically continue to follow a traditional path (simple random sampling or perhaps, stratified sampling) that fails to take advantage of significant advances in modeling for spatial data. In this paper, we consider sampling strategies for collecting environmental and biological samples that attempt to achieve approximately optimal spatial design performance using criteria we will describe below.

2. BACKGROUND AND LITERATURE REVIEW

A brief review of the history of spatial modeling for environmental health may be useful. Two broad paths have been followed. The first views the surface as a random realization of a spatial process above two-dimensional space. Measurements are taken at point-referenced (geo-coded) spatial locations. Inference involves fitting an explanatory process model using these measurements. In some cases, exposure levels are essentially stable and static modeling based upon single measurements at individual locations is the objective. In other cases, the locations are monitoring stations whence data collection is dynamic and a temporal component is added to the modeling to capture evolution of the contaminant surface over time. The literature on spatial and spatio-temporal process modeling in environmental health is substantial. Noteworthy examples for the static case include Le and Zidek (1992), Brown *et al*. (1994), Shaddick and Wakefield (2002), and Schmidt and Gelfand (2003). Examples in the dynamic setting include Guttorp *et al*. (1994), Huerta *et al*. (2004), and Sahu *et al*. (2005). Gelfand *et al*. (2005) provide a general dynamic modeling development for univariate and multivariate spatial data settings.

The second path has focused on areal partitions of the study region into, for example, census units, zip codes, or counties. Typically, counts of some adverse health outcome are aggregated to these units (usually for purposes of confidentiality). Environmental risk factors are supplied for these areal units to explain the counts. Spatially-structured random effects are introduced to provide spatial smoothing of the counts. Work here dates to Clayton and Kaldor (1987). See also Bernardinelli and Montomoli

(1992), Knorr-Held (2002), and Zhu *et al*. (2003). More flexible regression settings are discussed in Assunção (2003).

With regard to sampling for point-referenced data, we first note that optimal experimental design has a long statistical history. See the book of Pukelsheim (1993) for a review. The dominant path has focused on design for independent data collection. There is history for the case of correlated data dating to 1966 (Sacks and Ylvisaker, 1966). More recently, there has been attention directed at accommodating data with structured dependence. For spatial data, this has been expressed through random fields. See the review paper of Fedorov (1996) and the book of Müller (2001). Generally, designs are classified as either probability or model-based. The former includes widely-used simple random sampling without replacement. They tend to be robust in that they make no population assumptions regarding, for example, mean structure or dependence structure.

Model-based design has followed a regression model path or a random field model path. Under regression modeling with independent data, optimality is defined with regard to efficiency of the estimates of the regression coefficients. An optimization criterion that is a function of the design matrix is specified and then the 'best' design optimizes this criterion over all design matrices. Again, see Pukelsheim (1993) or Müller (2001) for details. This theory is not directly extensible to spatial design but approximately optimal solutions based upon information-theoretic measures have emerged, most notably the recursion in Brimkulov *et al*. (1986). (See Fedorov, 1996, in this regard.) This recursion is elaborated in Section 4 below. Its focus is exclusively on gaining information regarding the regression structure or model mean. A different type of approximation in the context of anisotropic dependence is proposed in recent work of Arbia and Lafratta (2002).

Model-based design, motivated by a random field specification, has been strongly advocated in Le and Zidek (1992) and Zidek *et al*. (2000) as well as references therein. The proposal is an entropy-based design where the selection of the next site to be added will be the one with the largest entropy where entropy can be viewed as uncertainty. Under a Gaussian field assumption, the criterion that emerges is the conditional variance of an observation at a new location, given the locations already selected. (This conditional variance depends only upon the previously selected locations but not on the data already collected at those locations.) The site with the most uncertainty is the one with the largest conditional variance given the selected sites. Extension to multivariate data at a location converts the criterion to a conditional covariance matrix. This approach has no interest in mean structure. In fact, quoting Zidek *et al*. (2000, p. 66), '[I]t avoids the need to specify objectives like parameter estimation.' Implementations have been in the area of network monitoring design and thus, initial preparation and operating costs are built into the adopted optimization criterion.

For a stationary Gaussian process with regression structure, two types of design questions can be asked: what is the optimal sampling design for prediction at an unobserved set of locations? What is the optimal sampling design for estimation of the parameters in the covariance function? Because prediction is often the primary use for the model, the first question has received much attention. See, for example, McBratney *et al*. (1981), Su and Cambanis (1993), Ritter (1996), and Zhu (2002). For the latter question, with a constant mean, the classical procedure for estimating the covariance structure is based upon the variogram. See, for example, Warrick and Myers (1987), Bogaert and Russo (1999), and Müller and Zimmerman (1999). Very recent work by Zhu and Stein (2005) focuses on designs based upon optimization using the likelihood. They suggest working with the Fisher information as a measure in the form of a ratio of determinants and implement the optimization using a simulated annealing algorithm.

Our approach is to also be model-based, working with the likelihood and focusing on the information matrix as well. As noted above, our perspective is primarily pragmatic. We conceive a

sampling setting in which we envision hundreds to thousand of sites being sampled and seek to make the required sampling design easily understood and computationally manageable for the practitioners who wish to implement it. We take as our design objectives learning about the mean structure as well as the covariance function, noting that these objectives are usually in conflict. We also introduce a further utility notion, providing an additional objective of sampling for say, large values (as with contaminant surfaces). We consider the situation where we already have a partial sample and we wish to augment the available data. After clarifying that obtaining the optimal solution is a combinatorially complex computation problem, we consider three approaches toward achieving approximate optimization—sequential selection, block selection, and tuned stochastic search.

The format of this paper is as follows. In Section 3, we review the issues in design development in our setting. In Section 4, we formalize the information-based design criterion we work with. Section 5 proposes several approximately optimal sampling schemes, which can be implemented with the criterion. Section 6 takes up information gain and connection to the entropy criterion. Section 7 considers comparison among the proposed sampling approaches. Section 8 looks at utility-based modification of the design criterion. Section 9 offers a clarifying simulation illustration with Section 10 proposing an illustrative application. We conclude with a summary and suggestion of future directions.

3. AN OVERVIEW OF THE ISSUES

Our objective is, for a given study region, to develop an approximately optimal sampling strategy to learn about the spatial distribution of a contaminant across the region. Optimal design is intractable working with the continuum of locations so, as is customary, we presume that the region has been gridded (not necessarily a regular grid) to high resolution. For instance, in the context of sampling childhood blood lead levels, the tax parcel level (equivalently the residential property on the parcel) provides a natural discretization for sampling locations. In the ensuing development, we assume that the parcels can be viewed as points in the region but, ultimately, with regard to design, we have only a finite set of locations to select from.

We will work within the model-based framework for developing designs. The two types of criteria we might consider are

(1) An information criterion that arises from the regression perspective (Section 4) but incorporates learning about strength of spatial dependence as well as the regression component.
(2) An entropy criterion that focuses on uncertainty, yielding a conditional covariance. We emphasize the information criterion in this paper for reasons we elaborate in Subsection 6.2. However, we reserve Subsection 6.2 for some comparison.

We also note that in the multiparameter case (almost certainly the case of interest in applications), both criteria emerge as matrices. So, to achieve a single number summary for a design, we will have to summarize the resulting matrix either through a determinant or a (possibly weighted) trace.

We further assume that sampling is not *ab initio* or 'preposterior.' Rather, we assume that a collection of n locations have already been sampled and that we have this data available to us. Such collection may have been implemented by simple random sampling or perhaps, through *ad hoc* methods. If not, how should the initial set of n points be selected? A convenient approach referred to as space-filling designs, has been discussed in Nychka and Saltzman (1998). Such designs are based upon geometric measures of how well a given set of points covers the study region, independent of the

assumed covariance function. Such designs are not optimal but for an initial selection will work nearly as well as optimal ones. In any event, as the number of sampling sites grows, effects of the initial selection dissipate.

Based upon the data from these n sites, we can implement a preliminary fit of the model to obtain preliminary parameter estimates. This is crucial since our design criteria emerge as parametric functions. To evaluate a criterion for a given set of locations, we insert the parameter estimates into the function as well as the locations. We recognize that this fails to account for the uncertainty in the parameter estimation and that averaging over a suitable distribution for these parameters would enable us to attach uncertainty to the criterion value. However, with interest in design rather than inference (which would come later, after all of the data collection) we adopt the pragmatic 'plug in' approach. It is also computationally much more convenient and avoids the need for prior specification at the time of sampling.

Thus the formal goal is, given the n locations already chosen and sampled, and given that we want to choose m additional sites to sample, how shall we choose these m locations? Even given an explicit, evaluable criterion and a finite collection of 'N' sites to choose from, obtaining the optimal choice is not a tractable problem. In our setting it would be referred to as an '$N - n$ choose m' combinatorially hard problem. So, we will have to consider approximate solutions to this problem. We examine three strategies: (i) sequential selection, (ii) block selection; and (iii) stochastic search (including a modified procedure).

4. THE INFORMATION CRITERION

In presenting the information criterion, rather than elaborating the formal optimal design machinery (as described, for instance, in Pukelsheim (1993) or in Müller (2001)), we offer an intuitive development built from the well-established Fisher information measure (see, e.g., Rao,1973; Cox and Hinkley, 1974). The Fisher information arises from expectation of second derivatives of the log likelihood. In the multiparameter case, it becomes the expectation of a matrix of mixed partial derivatives (the Hessian) associated with the log likelihood. Under normality and a linear mean form (in the coefficients) it emerges as a parametric function of the dependence structure. The matrix is reduced to a scalar criterion either through the trace or determinant.

More precisely, suppose we consider the widely used spatial model

$$Y(s_i) = \mu(s_i) + W(s_i) + \epsilon(s_i), \tag{1}$$

where $\boldsymbol{Y}(s_i)(i = 1, \ldots, n)$ are observations from a spatial process over a region D in $\mathbf{R}^2$ and $\mu(s_i)$ is the linear mean form, $\boldsymbol{X}(s_i)^{\mathrm{T}}\boldsymbol{\beta}$. $W(s_i)$ are realizations from a mean 0 spatial process (typically a stationary Gaussian process) and $\epsilon(s_i)$ are realizations from a pure error process with mean 0 and variance τ^2. W and ϵ are independent. Written in vector form $\boldsymbol{Y} = \boldsymbol{X\beta} + W + \epsilon$ where

$$\begin{pmatrix} W \\ \epsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2\mathbf{R}(\eta) & 0 \\ 0 & \tau^2\mathbf{I}_n \end{pmatrix}\right).$$

Here, $\mathbf{R}(\cdot)$ is the correlation matrix associated with the n locations and η indexes the parameters of the correlation function: for example, in the Matérn case (see, e.g., Banerjee *et al*., 2004), a smoothness parameter and a range parameter.

Let $\boldsymbol{\theta} = [\sigma^2, \eta, \tau^2]^{\mathrm{T}}$ with $\Sigma_\theta = \sigma^2 \mathbf{R}(\eta) + \tau^2 \mathbf{I}_n$. The log likelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_\theta| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\Sigma_\theta^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The score function $S(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ is

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\Sigma_\theta^{-1}\boldsymbol{X}$$

and the associated Hessian is

$$\mathbf{H}_\beta \ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = -\boldsymbol{X}^{\mathrm{T}}\Sigma_\theta^{-1}\boldsymbol{X}.$$

So, the expected information matrix for $\boldsymbol{\beta}$ is $\mathbf{I}(\boldsymbol{\beta}) = -\mathbf{E}(\mathbf{H}_\beta \ell(\boldsymbol{\beta}, \boldsymbol{\theta})) = \boldsymbol{X}^{\mathrm{T}}\Sigma_\theta^{-1}\boldsymbol{X}$.

The score function for the ith component of $\boldsymbol{\theta}$ is

$$\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \theta_i} = -\frac{1}{2}\mathrm{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\right) + \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

and the ijth entry of the associated Hessian matrix $\mathbf{H}_\theta \ell(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = {} & \frac{1}{2}\mathrm{tr}\left[\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_j}\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i} - \Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i \partial \theta_j}\right] + \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \\ & \left[-2\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_j}\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1} + \Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i \partial \theta_j}\Sigma_\theta^{-1}\right](\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).\end{aligned}$$

Hence, the ijth entry of the expected information matrix of $\boldsymbol{\theta}$ is

$$\mathbf{I}(\boldsymbol{\theta})_{ij} = -\mathbf{E}\left[\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] = \frac{1}{2}\mathrm{tr}\left[\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_j}\right]. \tag{2}$$

Finally, the expected information matrix for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ has the block diagonal form

$$\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{X}^{\mathrm{T}}\Sigma_\theta^{-1}\boldsymbol{X} & 0 \\ 0 & \left(\frac{1}{2}\mathrm{tr}\left[\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_j}\right]\right) \end{pmatrix}. \tag{3}$$

The block diagonal form in Equation (3) shows that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are *orthogonal* parameters (Cox and Reid, 1987). Informally, this means that an information criterion for design will 'separate' information regarding $\boldsymbol{\beta}$ from information regarding $\boldsymbol{\theta}$.

As it stands, Equation (3) is not a *criterion*. We need to reduce it to a univariate summary, which we will then seek to maximize. Such optimization will correspond to maximizing information gain, as we

detail in the next section. To achieve such reduction, we introduce a mapping from information matrices to scalars. Customary approaches work with either $\mathrm{tr}(\mathbf{I}(\boldsymbol{\beta},\boldsymbol{\theta}))$ or $|\mathbf{I}(\boldsymbol{\beta},\boldsymbol{\theta})|$. The former emerges as $\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\Sigma_{\theta}^{-1}\boldsymbol{X}) + \sum_i 1/2\mathrm{tr}\left[\Sigma_{\theta}^{-1}\partial\Sigma_{\theta}/\partial\theta_i\Sigma_{\theta}^{-1}\partial\Sigma_{\theta}/\partial\theta_i\right]$, the latter as $|\boldsymbol{X}^{\mathrm{T}}\Sigma_{\theta}^{-1}\boldsymbol{X}|\times |\mathrm{tr}\left[\Sigma_{\theta}^{-1}\partial\Sigma_{\theta}/\partial\theta_i\Sigma_{\theta}^{-1}\partial\Sigma_{\theta}/\partial\theta_j\right]2|$. In either case, we see the separation mentioned above. The forms also reveal that there can be tension between the component terms. That is, for a fixed m, the set of points which maximizes our information gain about $\boldsymbol{\beta}$ will be different from that for $\boldsymbol{\theta}$. Also, the trace criterion suggests the possibility of weighted components (see Section 9).

Here, and in the sequel, we work with the trace of the information matrix to provide our criterion rather than the determinant. The former is more intuitive in appreciating the components in the information gain but requires standardization of the covariates since it is not independent of the scale of the covariates (as the form $\boldsymbol{X}^{\mathrm{T}}\Sigma_{\theta}^{-1}\boldsymbol{X}$ reveals). The latter avoids that problem since any scaling emerges as a constant multiple of the determinant but at the expense of ease of interpretation.

Lastly, the response can be modeled on a suitably transformed scale in order to make the Gaussian assumption more comfortable. Moreover, in what follows we work with the foregoing modeling assumptions because they yield convenient computational expressions. We are not restricted to this setting; with additional computational effort, we can accommodate non-Gaussian data, for example, categorical outcomes or counts and/or non-linear means.

5. APPROACHES FOR APPROXIMATELY OPTIMAL DESIGN

To address the '$N-n$ choose m' combinatorially hard design problem noted at the end of Section 3, we consider three approximate solutions, sequential selection, block selection, and (modified) stochastic search. We consider them individually here though one could readily envision hybrid versions.

The sequential approach would require us to (i) identify, as the $(n+1)$st parcel, the one which provides the maximum increase in information, (ii) sample it and add its data to the data already collected, (iii) revise our current information, now based upon $n+1$ parcels, (iv) reorder the remaining parcels, and (v) select the $(n+2)$nd. In fact, we can make a modest compromise (which is appropriate for the way that the data collection would likely proceed) by sequentially ordering the parcels but only assuming we have data about the underlying process model from the first n locations. In this fashion, we can order the next m parcels to be selected. Then, if additional sampling were sought after these new m locations are sampled, we would refit the model and revise our knowledge about the model parameters in order to further sample.

The block selection approach would order all of the remaining parcels, given the n already selected and then choose the m parcels with the m largest values of the criterion. Evidently, it offers computational savings. Again, after these new m parcels were sampled, we would update our parameter estimates before further sampling. Hence, either scheme provides an ordering to all of the unsampled parcels. However, as we clarify below, these two approaches provide dramatically different sampling designs and, though the sequential scheme emerges as generally preferable, we can not assert that for any N, n, and m, it will always ensure greater information gain.

The third approach introduces stochasticity into the selection process. The most naive stochastic selection algorithm would choose m points at random and would be simple random sampling. Stochastic search is introduced if we make, say, b random selections, calculate the information gain for each, and adopt the one yielding the largest gain. Of course, the choice of b is unclear. The larger b is the closer we must get to the optimal design; however, computation cost increases linearly in b. Refinement of the stochastic search is possible. For instance, consider any location s which was

selected in at least one of the b searches. We can compute the average information gain for this location over all of the searches in which it was included. We could then propose to sample the m locations providing the largest average information gains.

We do note that, though it is not feasible to obtain the optimal design, the fact that we are dealing with a finite set of locations N does enable us to compute an upper bound on the information gain, that is, the information gain associated with sampling all remaining $N - n$ locations. Evidently, the information gain for choosing m points will tend to this bound as m increases. In fact, this raises an important theoretical point that we discuss briefly in Section 7. What can we say about $\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\theta})$ as N grows large? What can we say about $\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\theta})$ for fixed N as m increases? For the former, the key point is whether information tends to ∞ as $N \to \infty$ or remains bounded. For the latter, typically the information gain increases rapidly over smaller m with diminishing returns from there on. Hence, the upper bound not only provides a measure of what proportion of potential information we will gain from our sample of size m but also, if we see an 'elbow' in the information gain as a function of m, we might conclude that there will be little value in spending for additional sampling.

6. INFORMATION GAIN AND COMPARISON TO THE ENTROPY CRITERION

6.1. Calculation of the information gain

Returning to the model in Equation (1), recall that we seek to learn about the importance of the covariates (the $\boldsymbol{X}$s) in explaining the responses (the $\boldsymbol{Y}$s) as well as the nature and strength of spatial dependence. Assume the Gaussian process has stationary covariance function $\sigma^2\rho(s - s'; \phi)$.[1] Here, we assume $\rho \geq 0$ and that ρ strictly decreases from 1 to 0 as ϕ goes from 0 to ∞. A typical example is the so-called exponential covariance function with $\rho = \exp(-\phi||s - s'||)$. Also included are the powered exponential and Matérn families of covariance functions of which the exponential is a special case. We also assume, for the moment, that $\tau^2 = 0$, that is, there is no nugget.

It is a routine calculation to show that, if the $\boldsymbol{Y}$s all have a common mean say μ, up to a constant, the information in the sample about μ given spatial dependence measured by ϕ is the scalar $\mathbf{I}_n(\mu) = 1^{\mathrm{T}}\mathbf{R}_n^{-1}(\phi)1 \equiv \mathbf{A}_n$ where $\mathbf{R}_n(\phi)$ is the $n \times n$ matrix with (i,j) entry $\rho(s_i - s_j; \phi)$. (So, we can ignore the unknown σ^2 in comparing designs.)

Despite its innocuous form, general behavior for $\mathbf{A}_n$ is not easy to prove except in very special cases (see Section 7). However, we can explicitly compute the information gain in sampling location s_0. We have

$$\mathbf{A}_{n+1} - \mathbf{A}_n = \frac{(1 - 1^{\mathrm{T}}\mathbf{R}_n^{-1}(\phi)\mathbf{R}_{n0}(\phi))^2}{1 - \mathbf{R}_{n0}^{\mathrm{T}}(\phi)\mathbf{R}_n^{-1}(\phi)\mathbf{R}_{n0}(\phi)}. \tag{4}$$

In this expression, $\mathbf{R}_{n0}(\phi)$ is an $n \times 1$ vector with ith entry $\rho(s_i - s_0; \phi)$. So, the s_0 that maximizes this difference is the location that maximizes information gain. The maximization is easy to carry out since we only have a finite number of sites and since $\mathbf{R}_{n0}(\phi)$ changes with s_0 but $\mathbf{R}_n^{-1}(\phi)$ does not. In fact, we suggest the creation of a GIS display in the form of a choropleth map or a contour plot to reveal where in the region information gain is high and where it is low. (See the illustrative example in Section 9.) Evaluation of the criterion requires knowing the covariance function, that is, requires

[1] In fact, this is not required but does simplify the ensuing presentation.

estimating ϕ. As discussed above, this will be done using the n data points already collected. That is, the initial data provides our starting knowledge regarding spatial structure. As we collect additional data, we use it to revise our learning about this structure.

As noted above, $\mathbf{A}_n$ calculates the information in the sample about the mean μ. There is also information in the sample about σ^2 and ϕ. In particular, $\mathbf{I}(\mu, \sigma^2, \phi)$, as a special case of Equation (3), takes the form, for sample size n,

$$\mathbf{I}_n(\mu, \sigma^2, \phi) = \begin{pmatrix} \mathbf{A}_n/\sigma^2 & 0 & 0 \\ 0 & n/\sigma^2 & \mathbf{B}_n/\sigma \\ 0 & \mathbf{B}_n/\sigma & \mathbf{C}_n \end{pmatrix}, \tag{5}$$

where $\mathbf{A}_n$ is as above, $\mathbf{B}_n = \mathrm{tr}(\mathbf{R}_n^{-1}\partial\mathbf{R}_n/\partial\phi)$ and $\mathbf{C}_n = \mathrm{tr}(\mathbf{R}_n^{-1}\partial\mathbf{R}_n/\partial\phi\mathbf{R}_n^{-1}\partial\mathbf{R}_n/\partial\phi)2$. Hence $|\mathbf{I}_n(\mu, \sigma^2, \phi)|$ has the simple form $\sigma^{-4}\mathbf{A}_n(n\mathbf{C}_n - \mathbf{B}_n^2)$ explicitly revealing the separation in information contributions. Suppose that our interest focuses on the information gain for both μ and ϕ (i.e., we pretend that σ^2 is known). We can simplify (5) to

$$\mathbf{I}_n(\mu, \phi) = \frac{1}{\sigma^2}\begin{pmatrix} \mathbf{A}_n & 0 \\ 0 & \sigma^2\mathbf{C}_n \end{pmatrix}. \tag{6}$$

Taking the trace of this matrix revises the criterion to $\mathbf{A}_n/\sigma^2 + \mathbf{C}_n =$

$$1^{\mathrm{T}}\mathbf{R}_n^{-1}(\phi)1/\sigma^2 + \frac{1}{2}\mathrm{tr}\left\{\mathbf{R}_n^{-1}(\phi)\frac{\partial\mathbf{R}_n(\phi)}{\partial\phi}\mathbf{R}_n^{-1}(\phi)\frac{\partial\mathbf{R}_n(\phi)}{\partial\phi}\right\}. \tag{7}$$

Evaluation of Equation (7) requires estimating σ^2 and ϕ. Again, this will be done using the n data points already collected. Using convenient matrix identities (e.g., Harville, 1997), computational methods for the rapid calculation of the analogue of Equation (4) are available; we omit details. In fact, it may be of interest to compare the approximately optimal designs for just the first term in Equation (7) or just the second term in Equation (7). (see Section 9.) However, for the remainder of this section we omit the contribution of ϕ to the information gain.

For each site we typically have available covariate information, for example, for tax parcels, the age of the house on the parcel might be assumed to provide explanation regarding the presence of biologically available lead at the location. Suppose in subsequent analysis, once the data is collected, we anticipate using such information in the mean specification, say in the form of a linear regression, $\beta_0 + \beta_1 X(s)$ where $X(s)$ is the age associated with parcel s. Then, we might seek to choose the parcels to maximize the information about the linear function in age, that is, in β_0 and β_1. (Note that this is not the same objective as choosing sites to encourage $Y(s)$ to be large, see Section 8 below.)

More generally, with a $p \times 1$ vector of covariates $X(s)$, including the intercept, we obtain a $p \times p$ information matrix (the upper left matrix in Equation (3)). Using the trace, we now obtain $\sum_{l=1}^{p} \boldsymbol{X}_l^{\mathrm{T}}\mathbf{R}_n^{-1}(\phi)\boldsymbol{X}_l$ where $\boldsymbol{X}_l$ is $n \times 1$ with ith entry $\boldsymbol{X}_l(s_i)$. Again we can compute the information gain explicitly in selecting parcel s_0. In fact, we obtain

$$\frac{\sum_{l=1}^{p}(\boldsymbol{X}_l(s_0) - \boldsymbol{X}_l^{\mathrm{T}}\mathbf{R}_n^{-1}(\phi)\mathbf{R}_{n0}(\phi))^2}{1 - \mathbf{R}_{n0}^{\mathrm{T}}(\phi)\mathbf{R}_n^{-1}(\phi)\mathbf{R}_{n0}(\phi)}.$$

This is the recursion of Brimkulov *et al*. (1986).

6.2. Comparison of the information criterion and the entropy criterion

From Section 1, the entropy criterion is phrased in terms of extent of uncertainty and is motivated by work in pollution-monitoring network design as summarized in Zidek *et al*. (2000). A scalar arises in the univariate case, the determinant of a matrix in the multivariate case. Again, with normally distributed, dependent data both the scalar and the determinant will be parametric functions of the dependence structure. In the design setting it is intuitively easiest to interpret entropy as uncertainty. Sites with high entropy, given those that we have already sampled, would be desirable choices to select. That is, from the remaining sites, we would seek to learn about those for which we are most uncertain. Hence, the criterion computes the entropy given the current set of sites and adds next the site with the largest conditional entropy. With the assumptions and notation above, it is straightforward to show that the conditional entropy associated with site s_0 is $1 - \mathbf{R}_{n0}^{\mathrm{T}}(\phi)\mathbf{R}_n^{-1}(\phi)\mathbf{R}_{n0}(\phi)$ (Zidek *et al.*, 2000). As a conditional variance, this quantity is obviously non-negative and so, we choose s_0 to maximize this. Computation is straightforward. A choropleth map or contour plot of the values of this criterion over the collection of parcels would provide a useful display.

It is interesting to note that the entropy criterion is the denominator of the information criterion. This appears paradoxical since we are proposing to maximize both criteria. In fact, the square in the numerator of the information criterion offsets the denominator to remove the paradox. We can clarify by looking at the $n = 1$ case. The information criterion becomes $(1 - \rho)/(1 + \rho)$ while the entropy criterion becomes $1 - \rho^2$. Both decrease from 1 to 0 as ρ increases from 0 to 1. However, the functions are quite different; for instance, the former is convex while the latter is concave.

The entropy criterion can be extended to accommodate pure error as well, replacing $\mathbf{R}_n(\phi)$ with $\sigma^2\mathbf{R}_n(\phi) + \tau^2\mathbf{I}_n$ as in Section 4. The resulting form for the criterion is $\sigma^2 + \tau^2 - \sigma^4$ $\mathbf{R}_{n0}(\phi)^{\mathrm{T}}(\sigma^2\mathbf{R}_n(\phi) + \tau^2\mathbf{I}_n)^{-1}\mathbf{R}_{n0}(\phi)$. The criterion can also be extended to multivariate measurements in the form of the determinant of the conditional covariance matrix associated with $\boldsymbol{Y}(s_0)$. With a separable specification for the error structure, an argument similar to that for the information criterion enables us to use the same entropy criterion as above.

Finally, the criterion would not be affected by the introduction of covariate information for each site. The entropy measure focuses only on uncertainty arising from spatial structure. The conditional variance is not affected by the mean specification. As Zidek *et al*. (2000) note, the criterion avoids issues like parameter estimation and hypothesis testing. In our context, we would not view this as advantageous since we want to learn about the nature of the regression relationship between the level of the response and the proposed explanatory variables. So, for our purposes, the information criterion emerges as preferred. In particular, it will use the available $\boldsymbol{X}(s)$ vectors in the determination of the selection order.

7. SOME TECHNICAL REMARKS

Remark 1: Consider the model $Y(s) = \mu + W(s)$ and suppose that n locations $s_1, \ldots, s_n$ have already been sampled. Expression (4) shows that the information gain at s_0, $\Delta\mathbf{I}_{s_0}(\mu) > 0$ for a new s_0. In fact, $\Delta\mathbf{I}_{s_0}(\mu) \to 0$ when $||s_0 - s_i|| \to 0$, where $s_i (i = 1, \ldots, n)$ is any of the n samples and $||s_0 - s_i||$ is the Euclidean distance between s_0 and s_i. So, since $\mathbf{A}_1 = 1$, we have $\mathbf{A}_n > 1$ and $\mathbf{A}_n$ increases in n. If $\mathbf{R}_n \to \mathbf{I}_{n\times n}$ (the identity matrix), then $\mathbf{A}_n \to n$. Is $\mathbf{A}_n \leq n$? Since we showed in the previous section that $\mathbf{A}_2 - \mathbf{A}_1 = (1 - \rho)/(1 + \rho)$, if $\rho < 0, \mathbf{A}_2 > 2$ and, in fact, $\mathbf{A}_2 \to \infty$ as $\rho \to -1$.

Remark 2: Assuming that all the covariance parameters are given, the general behavior of $1^{\mathrm{T}}\mathbf{R}^{-1}(s_1, \ldots, s_n)1$ is surprisingly difficult to investigate. Results depend upon the form of ρ and

the nature of the asymptotics. For example, with a fixed region (so-called *infill* asymptotics; see Stein, 1999) and a separable covariance function that is a product of one-dimensional exponential covariance functions, we can compute $\mathbf{A}_n$ explicitly as well as its limit which is finite. More generally, for customary ρ that are isotropic, positive, and strictly decreasing, reaching 0 as distance reaches ∞, we can show empirically that $\mathbf{A}_n$ is bounded. If we allow the size of the region to grow as n grows, then the relative rates of growth determine the behavior of $\mathbf{A}_n$. Detailed discussion, including the foregoing results, is presented in Xia *et al*. (2005).

Remark 3: If one were to think in terms of choosing a distribution to randomly sample the locations from, intuition might suggest that the uniform distribution produces the maximum expected information. In fact, if sampling is for $s \in D$, $\mathbf{E}[1^{\mathrm{T}}\mathbf{R}^{-1}(s_1,\ldots,s_n)1|s_i \sim \mathrm{unif}(D)]$ will not maximize $\mathbf{E}[1^{\mathrm{T}}\mathbf{R}^{-1}(s_1,\ldots,s_n)1|s_i \sim f(D)]$ for all distributions f over D. Intuitively, appropriate systematic selection of points will provide greater information than the average under random selection. Consider the following simple example. Suppose we sample four points (s_1, s_2, s_3, s_4) uniformly on [0,1]. With the exponential correlation function $\mathbf{R}(s_i, s_j) = \mathrm{e}^{-7|s_i - s_j|}$, we can obtain, by Monte Carlo integration, $\mathbf{E}\{1^{\mathrm{T}}\mathbf{R}^{-1}(s_1,\ldots,s_4)1|s_i \sim \mathrm{unif}(0,1)\} = 2.52$ (Figure 1 shows the density function for the information in this case). However, if we choose $(s_1, s_2, s_3, s_4) = (0, 0.33, 0.67, 1)$, $1^{\mathrm{T}}\mathbf{R}^{-1}(s_1,\ldots,s_4)1 = 3.47$. Hence, a non-degenerate distribution that is not far from this *degenerate* choice will achieve a larger expectation than under uniform selection.

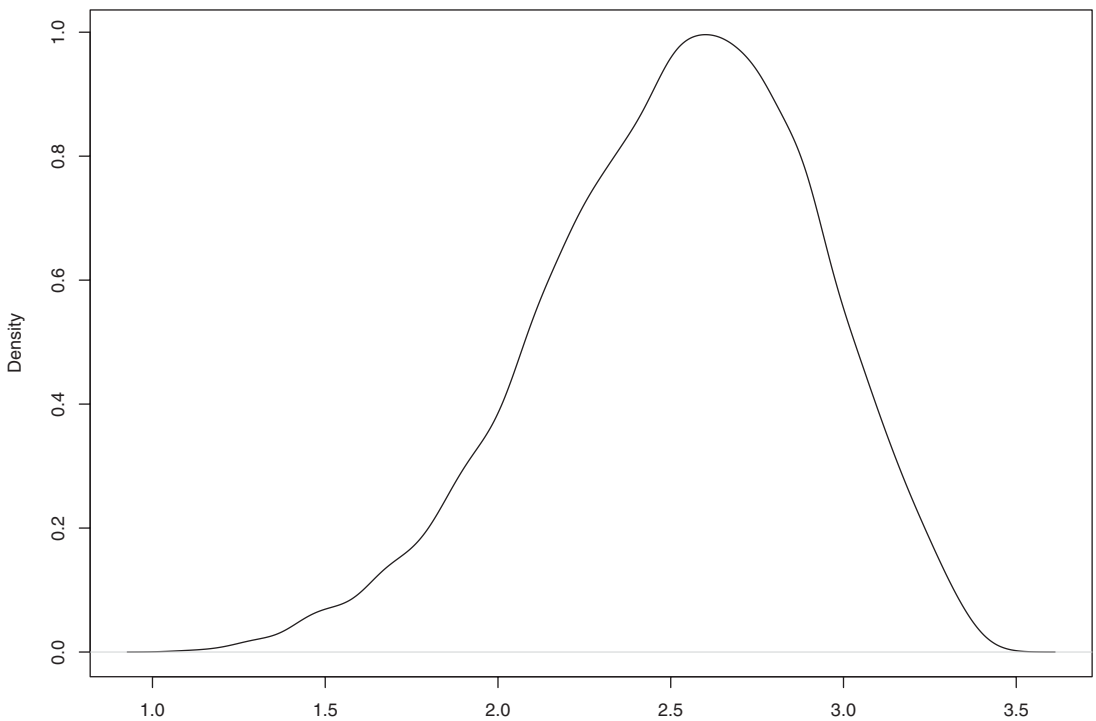

Figure 1. Density of $1^{\mathrm{T}}\mathbf{R}^{-1}1$ given $s_1,\ldots,s_4 \sim \mathrm{unif}(0,1)$

Remark 4: With infill asymptotics, if we return to $\mathbf{I}_n(\mu, \sigma^2, \phi)$ given in Equation (5), for customary ρ that are isotropic, positive and strictly decreasing, reaching 0 as distance reaches ∞, we can show empirically that $\mathbf{B}_n$ and $\mathbf{C}_n$ are unbounded. Again, see Xia *et al*. (2005).

Remark 5: In considering say, $\mathbf{I}_n(\mu, \sigma^2, \phi)$, the trace cumulates what we would define as the conditional information, for example, the information in the sample about ϕ given μ and σ^2 are known. We could also calculate *unconditional* information. We can show that the sum of the reciprocals of the diagonal elements of $\mathbf{I}_n^{-1}(\mu, \sigma^2, \phi)$ cumulates this unconditional information. Furthermore, the asymptotic behavior of unconditional information need not agree with that of conditional information.

Remark 6: We hope that it is clear that the sequential approach need not produce the optimal choice of m points. The useful analogy here is to variable selection in multivariate linear regression. A forward stepwise procedure is not guaranteed to produce the subset of variables of a fixed size which maximizes $\mathbf{R}^2$.

Remark 7: As an example to illustrate a case where sequential design will be worse than block design, suppose we have s_1 at the origin. We want to select three additional points to learn about the mean from $s_2, \ldots, s_5$ as shown in Figure 2. The block design will select (s_1, s_4, s_2, s_5) while the

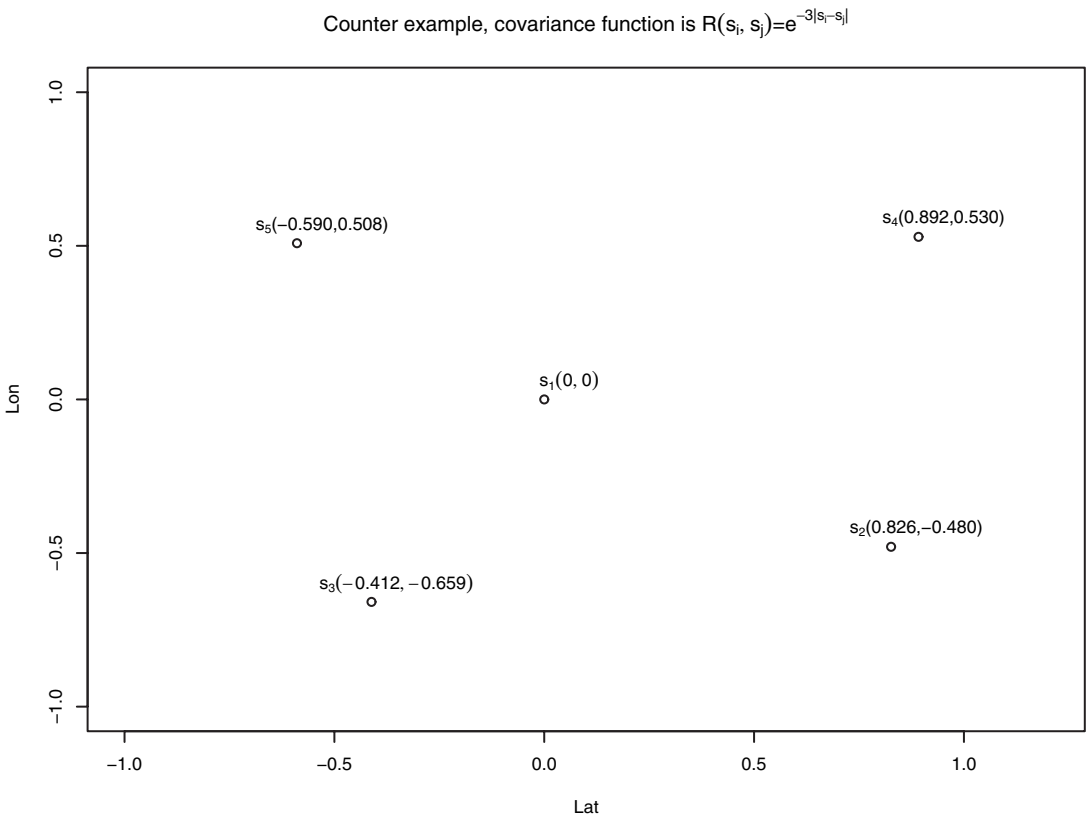

Figure 2. Sequential selection versus block selection

sequential design will select (s_1, s_4, s_3, s_2). The corresponding sequence of information values is (1, 1.915, 2.728, 3.539) and (1, 1.915, 2.736, 3.521), respectively; the four points selected by the block design produce greater information than those selected by the sequential design.

Remark 8: We conclude with an illustrative comparison among the sequential scheme, the block scheme, the stochastic search scheme ($b = 500$) and the refined stochastic search scheme (again, $b = 500$). In Figure 3 (which arises from the simulation illustration in Section 9), we plot the information growth for the intercept. The information for the intercept is bounded as noted in Remark 1 and the upper bound is given. The sequential design scheme is clearly the best, as it will be generally except for pathological examples such as in Remark 7. With 40 sites already sampled and 960 that could still potentially be sampled, more than 95 per cent of the upper bound is achieved with only 20 additional observations. For the block scheme and the modified stochastic search schemes roughly 70 additional observations are needed to do as well. The inferior performance of the simple stochastic search scheme is evident. We also plotted (Figure 4) the information growth for $\mathbf{I}(\phi)$, calculated through the lower right entry in Equation (3). Note the striking difference in the information scales between Figures 3 and 4. Also, we see that, with σ^2 fixed (known), information growth for ϕ is not bounded. See, for example, Xia *et al.* (2005) in this regard.

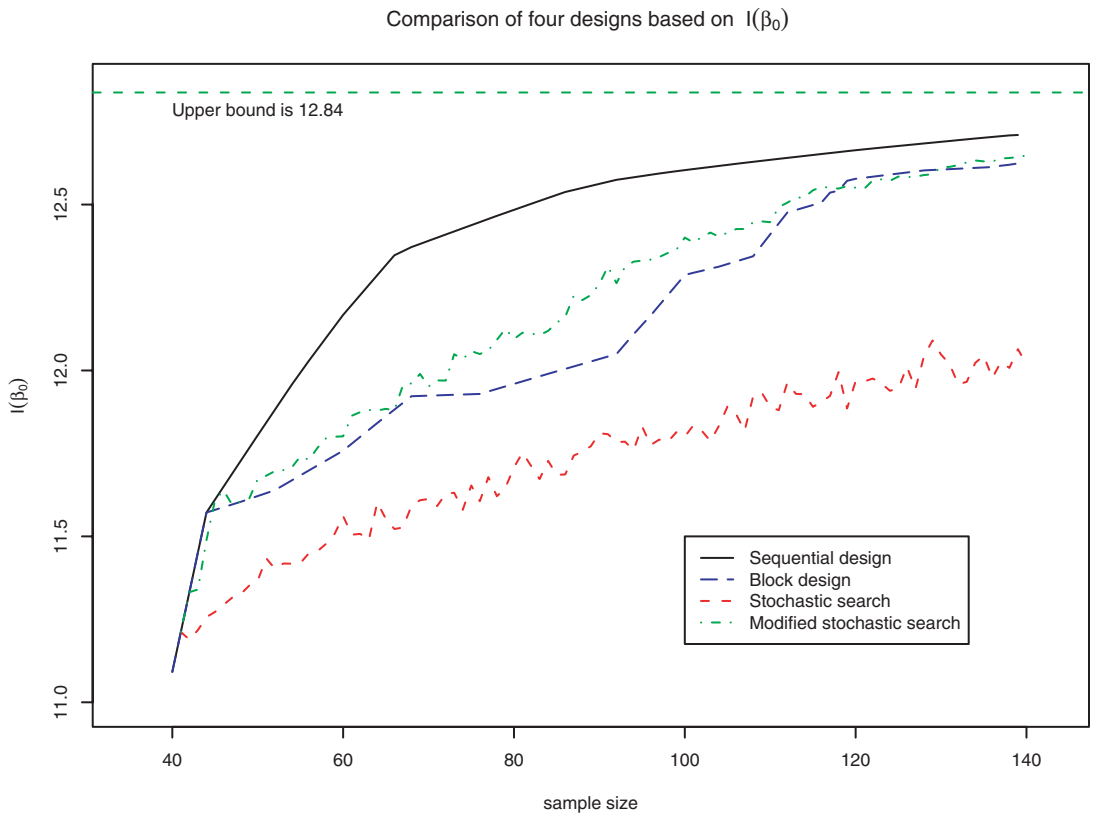

Figure 3. Information ($\mathbf{I}(\beta_0)$) growth in sample size for the four sampling schemes

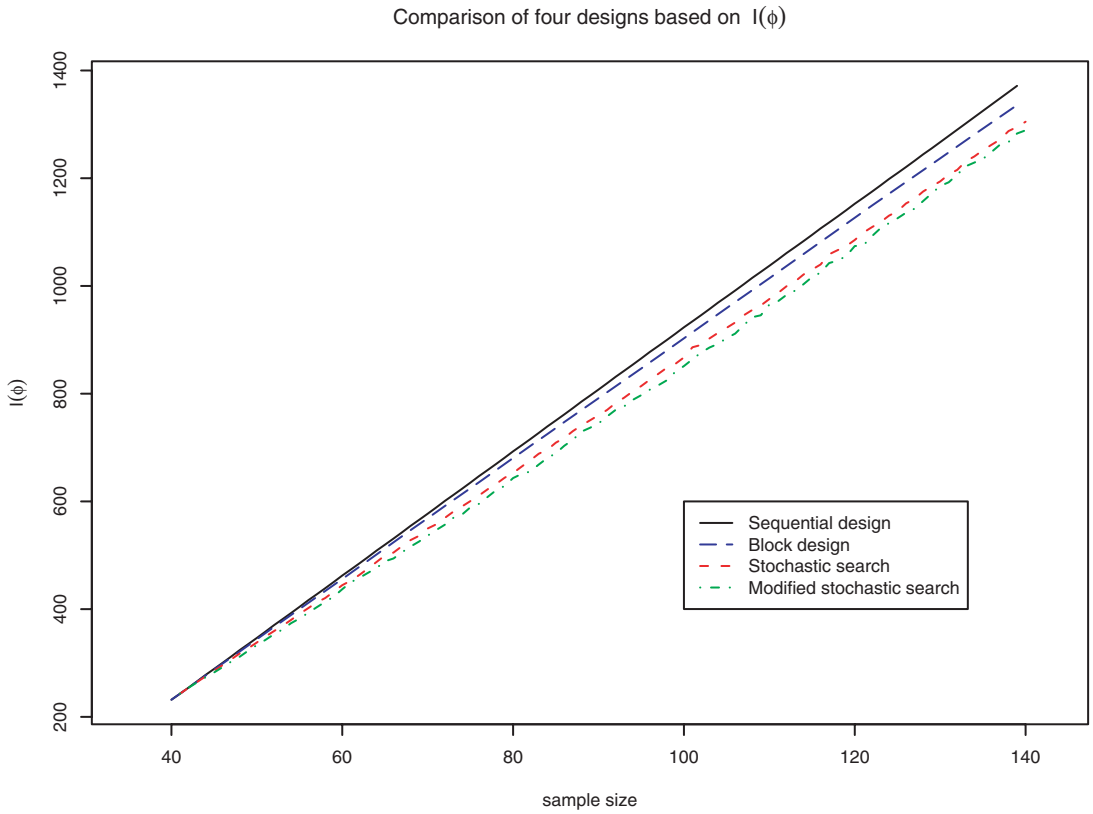

Figure 4. Information ($\mathbf{I}(\phi)$) growth in sample size for the four sampling schemes

8. MODIFIED UTILITY SPATIAL DESIGNS

The goal here is to propose the use of overlays of either (estimated) mean response or covariate data layers to achieve specific objectives, for example, to separate essentially equivalent locations under the foregoing criteria or to modify utility for point selection resulting in revised approximate optimization for the spatial sampling design.

For instance, the goal may be to learn about the regression relationship but this does not imply selection of sites where the response is expected to be high, for example, high levels of biologically available lead or of arsenic contamination. One way to achieve this is a model-based strategy, obtaining the estimated spatial surface based upon the data collected thus far, that is, based upon $Y(s_1), Y(s_2), \ldots, Y(s_n)$. Overlay of this surface on the selection surface will reveal parcels where both layers achieve high values in order to determine selection. Alternatively, we could multiply the surfaces to upweight/downweight the selection surface. One might also work not with the fitted model layer but, instead, a different data layer, perhaps external to those used in the model fitting. Such layers might reflect established geographic gradients with regard to say, the contaminant or distance from a site that is a known source for high contaminant levels.

This strategy would also address the matter of locations having essentially equivalent values under the criterion. They can be distinguished by using the second weighting layer, yielding a weighted

criterion. For instance, one could upweight parcels that are expected to exhibit high levels of the contaminant being sampled.

A somewhat different objective that could be used to distinguish parcels which are essentially equivalent under the information criterion would be to work with demographic data layers. In this case, the second objective would be to oversample parcels with certain demographic features, for example, in low socio-economic status areas or high crime rate areas. A spatial surface reflecting such a layer would be created. Again, overlaying or multiplication provides upweighting or downweighting of the selection surface. Ultimately, the issue is one of utility for the data collection. If we seek to learn not only about the exposure surface but also to achieve certain expected features in our samples, then we need to specify a utility function that reflects this objective.

9. COMPUTATIONAL ISSUES AND A SIMULATION ILLUSTRATION

In providing a simulation illustration, we focus on sequential design and block design to select an additional collection of m parcels from $N - n$ parcels given n have already been selected. We adopt the model $Y(s) = \boldsymbol{X}(s)^{\mathrm{T}}\beta + W(s) + \epsilon(s)$, and work with $\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\theta})$ as in Equation (3). For the sequential design we do not update the parameter estimates after each new location is selected. We only use the parameter estimation based upon the original n samples.

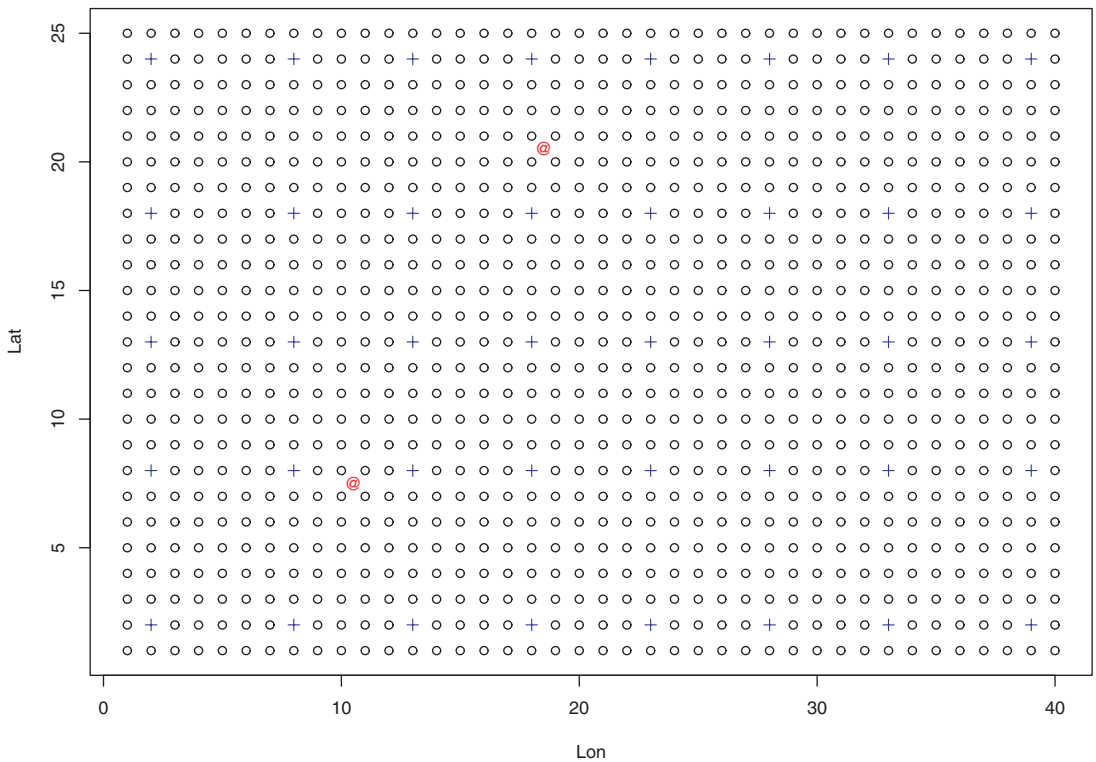

Figure 5. Study region

We generalize the trace of $\mathbf{I}(\beta, \theta)$ to define $\mathbf{I}(\beta) = \sum_{i=1}^{p} \omega_i \mathbf{I}(\beta_i)$ for vector $\boldsymbol{\beta}$ of length p, $\mathbf{I}(\theta) = \sum_{j=1}^{q} \nu_j \mathbf{I}(\theta_j)$ for vector $\boldsymbol{\theta}$ of length q and finally, the combined information as $\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{p} \omega_i \mathbf{I}(\beta_i) + \sum_{j=1}^{q} \nu_j \mathbf{I}(\theta_j)$. The weights allow us to rescale the components of the trace to reflect the fact that the information is affected by scale. For example, the information value would change if we convert distance from, say, kilometers to miles. Also the scale of information values for mean structure parameters $\boldsymbol{\beta}$ and for dependence structure parameters $\boldsymbol{\theta}$ can be very different (recall Figures 3 and 4). However, here we do not pursue standardizations (choices of v's and w's) since, as in the previous section, they too reflect utility for the design. Instead, in the illustration below, we show approximately optimal designs for $\boldsymbol{\beta}$, for $\boldsymbol{\theta}$ and for utility-weighted versions of these following Section 8. Furthermore, motivated by Figure 2, we work only with the sequential sampling scheme.

In particular, we turn to a simulation example where we conduct the spatial design based upon a grid of 40×25 parcels as Figure 5 shows.

We first sampled 40 of these parcels according to, for example, space-filling design at locations indicated by $+$ on the grid. We generated a random realization of a Gaussian process of the form $Y(s) = \beta_0 + \beta_1 X_1(s) + \beta_2 X_2(s) + W(s)$, ignoring the pure error term $\epsilon(s)$, for convenience. $X_1(s)$ denotes the distance of location s from a pollution source located at $(10.5, 7.5)$ while $X_2(s)$ denotes the distance of s from a different pollution source located at $(18.5, 20.5)$ (these are indicated by @ on the grid). The true $\beta_0 = 2$, the true $\beta_1 = 0.5$, and the true $\beta_2 = 1$. The spatial variability σ^2 is set to 1. We use the exponential covariance function with decay parameter $\phi = 0.2$, resulting in a spatial range of 31.75 per cent of the maximum distance in the region. Figure 6 is a three-dimensional perspective plot of the true mean surface.

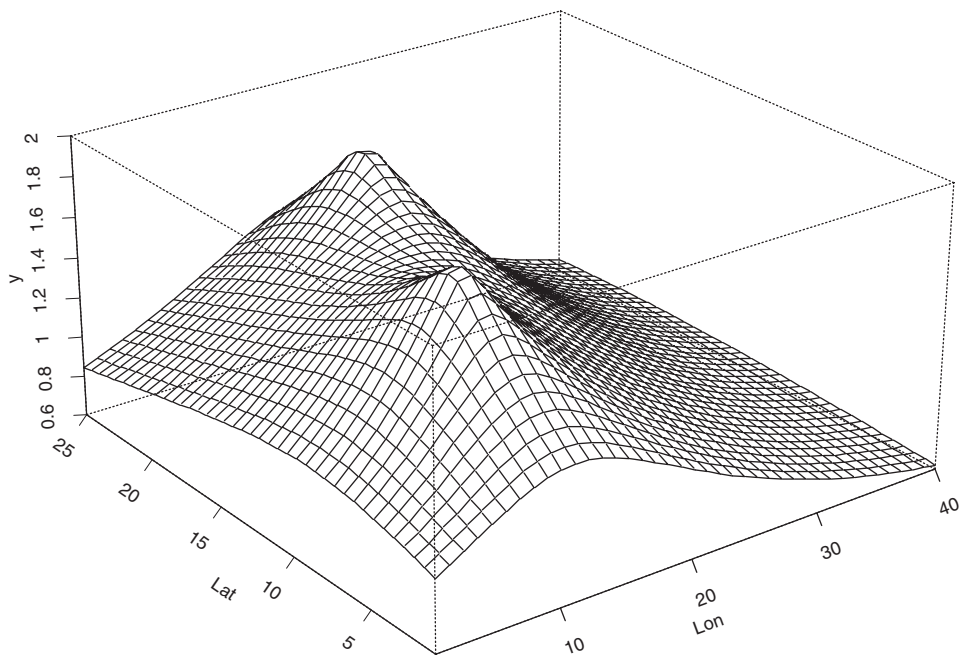

Figure 6. Mean surface

Figure 7 shows 100 selected locations(indicated by "•") based on the information gain of β_0. Though the selection sequence is not numbered (the figure becomes too cluttered if we do), adding 'less dependent', that is, 'most isolated' locations will most increase information. Notice also the striking edge effects which are inherent in sampling from a bounded region.

Figure 8 provides 100 selected points based on the information gain of $\beta_0, \beta_1, \beta_2$. It can be seen that, in addition to isolated locations, we also choose locations that are clustered around the two pollution sources. In this case, the information gain by adding a particular point depends on the distance from that point to the pollution sources as well as the dependence between that point and all the existing samples. Again, edge effects are strong.

Figure 9 shows the design that results from the criterion which attempts to maximize $\mathbf{I}(\phi)$, the gain in information about the spatial dependence. The choice of points is dramatically different from that for $\mathbf{I}(\beta_0)$. To learn about decay in spatial association, we need points near to each other. Edge effects are not an issue here.

Figures 10 and 11 provide the analogues of Figures 7 and 8 using the design points based upon a weighted criterion, in particular, weighted by the estimated mean at each location. Note that in the present case, Figure 10 changes dramatically from Figure 7 while Figure 11 is nearly the same as Figure 8.

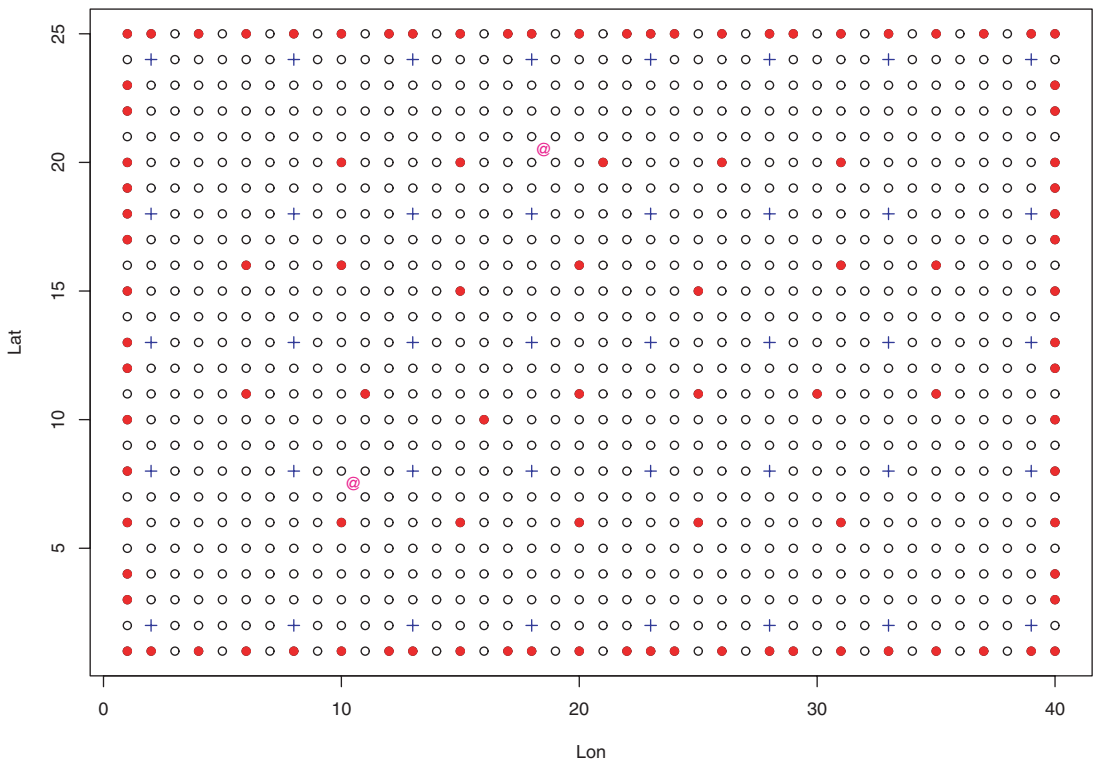

Figure 7. Sequential design based on $\mathbf{I}(\beta_0)$

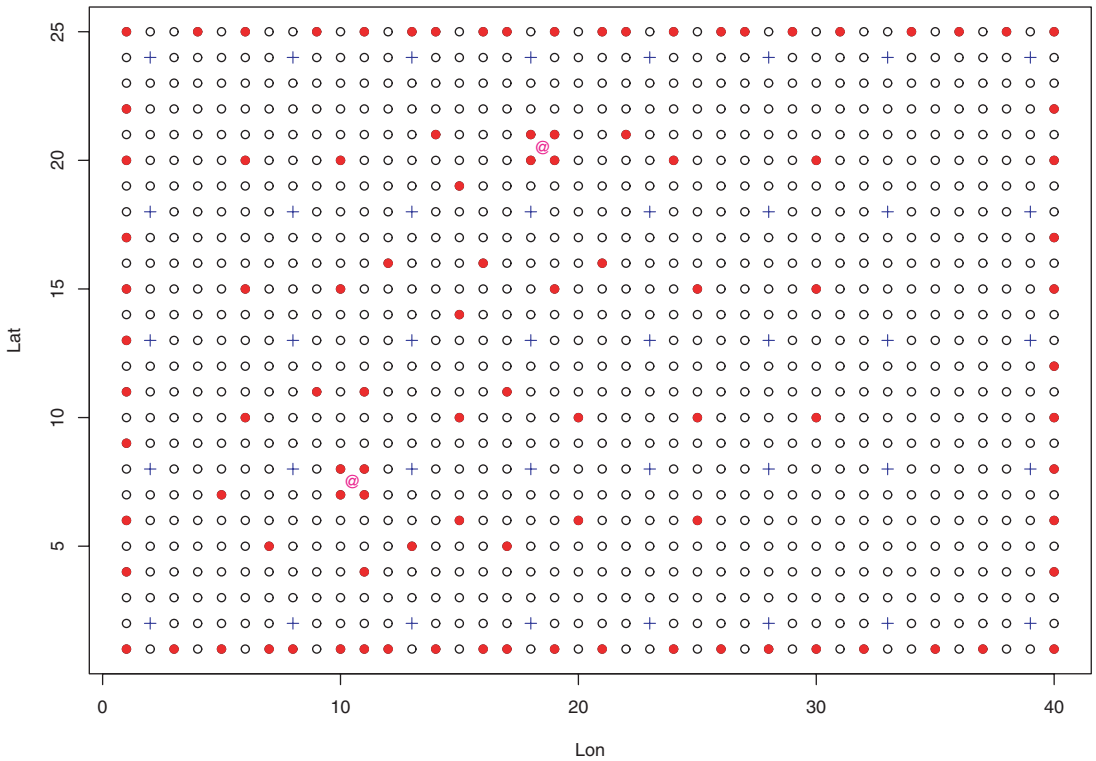

Figure 8. Sequential design based on $\mathbf{I}(\beta_0, \beta_1, \beta_2)$

10. A PROSPECTIVE REAL APPLICATION

In 1984, a Union Carbide plant in Bhopal, India released methyl isocyanate into the air at levels high enough to kill several thousand people in the immediate surrounding area. Not long after, Union Carbide's sister plant in South Charleston, West Virginia also experienced a significant release of acetone and mesityl oxide. Concerned for their safety, both industrial workers and local communities called for freely available information on the chemicals being used in and released from industrial facilities. In response to strong public demand, in 1986, the United States enacted the Emergency Planning and Community Right to Know Act.

Among other things, under Section 313 the Act established the Toxics Releases Inventory (TRI). In its original form, TRI required all businesses in Standard Industrial Classification (SIC) codes 20–39 that employed ten or more employees and released into the air, water, or ground either 10 000 pounds or more of any one of the 350 chemicals on the TRI list or 25 000 pounds or more of any combination of the 350 chemicals to report this information to the USEPA. SIC codes 20–39 cover the following industries: food, tobacco, textiles, apparel, lumber and wood, furniture, paper, printing and publishing, chemical, petroleum and coal, rubber and plastics, leathers, stone clay and glass, primary metals, machinery, electrical and electronic equipment, transportation equipment, instruments, and miscellaneous manufacturing. The TRI has subsequently been expanded to include metal mining, coal mining, coal- and oil-fired electric utilities, hazardous waste treatment and disposal facilities, chemicals and

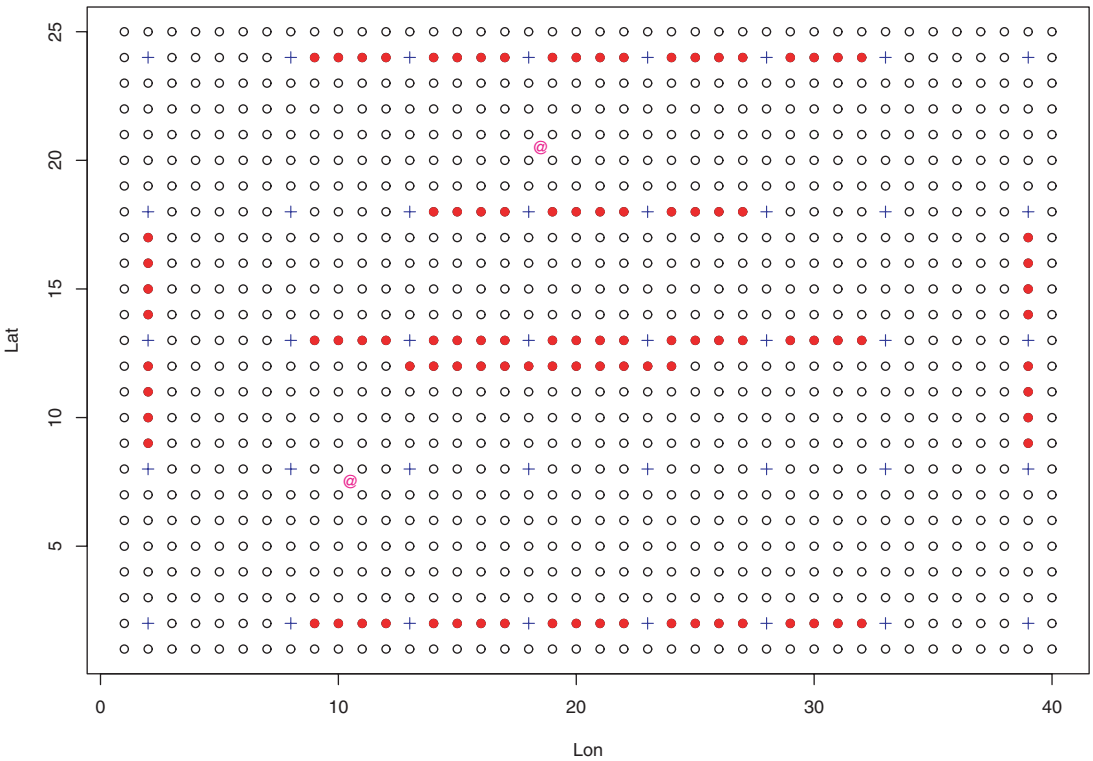

Figure 9. Sequential design based on $\mathbf{I}(\phi)$

allied products wholesale distributors, petroleum bulk plants and terminals and solvent recovery services, reflecting a total of 667 chemicals. In addition, the USEPA has recently reduced the reporting threshold for several chemicals that are considered either persistent or especially toxic, including hexachlorobenzene, mercury, and lead.

Reconsideration of the TRI reporting requirements includes questions regarding: (1) whether smaller facilities (fewer than 10 employees or lower chemical use levels) in TRI-reporting SIC codes should be required to report their emissions; (2) whether additional SIC codes should be required to report; (3) whether additional chemicals should be added to the TRI list; (4) whether reporting thresholds should be lowered on particular compounds (as was done for hexachlorobenzene, mercury, and lead; (5) whether facilities should be required to report both use and emissions; and (6) whether facilities that previously reported, but do not report currently, should be required to provide an explanation for this change in status. All of these policy questions are substantially hampered by the lack of systematic data on ambient levels of air toxics. Given the paucity of existing data and the cost of collecting new data, an optimized method for sampling design is essential. Take, for example, question (1) above regarding whether smaller facilities should be required to report their emissions. Previous research where emissions are imputed to smaller facilities indicates that including smaller facilities has a substantial impact on the spatial distribution of modeled ambient air concentrations of contaminants (Dolinoy and Miranda, 2004). This work, however, necessarily relies on model-based estimates of ambient levels that result from dispersion models. Alternatively, the facilities that already

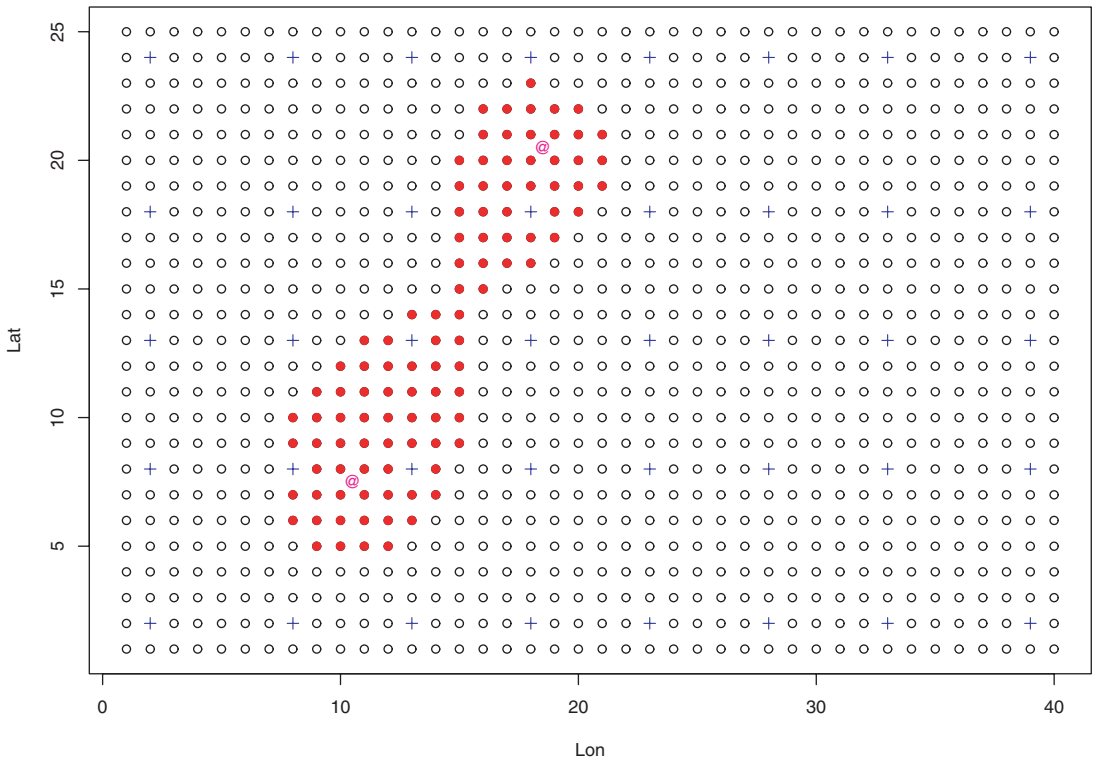

Figure 10. Sequential design based on $\boldsymbol{Y}^{*}\mathbf{I}(\beta_0)$

report to TRI would be analogous to the two point sources delineated in the simulation presented here. The smaller facilities represent known point sources with unknown emission levels. Since the emissions from the smaller facilities are unknown, they are not available for considering the question of whether these smaller facilities should be required to report. Determining how important the smaller facilities are to ambient concentrations can be much more efficiently accomplished through optimization of sampling design.

The optimized sampling design approach described can incorporate multiple sources of emissions and multiple chemicals emitted. The sampling design can also be shaped to specifically assess exposures to specific sub-populations whose geographic distribution can be characterized. Thus the approach holds great potential for helping scientists, agencies, and communities understand the distribution of TRI chemicals released into the environment.

11. DISCUSSION AND EXTENSIONS

We have considered approximately optimal environmental exposure sampling design for the setting where we expect intensive one-time sampling rather than sparse continuous monitoring. We have adopted information-based performance criteria and suggested a sequential implementation. We have shown that such a strategy is straightforward to implement with computational demand that is not

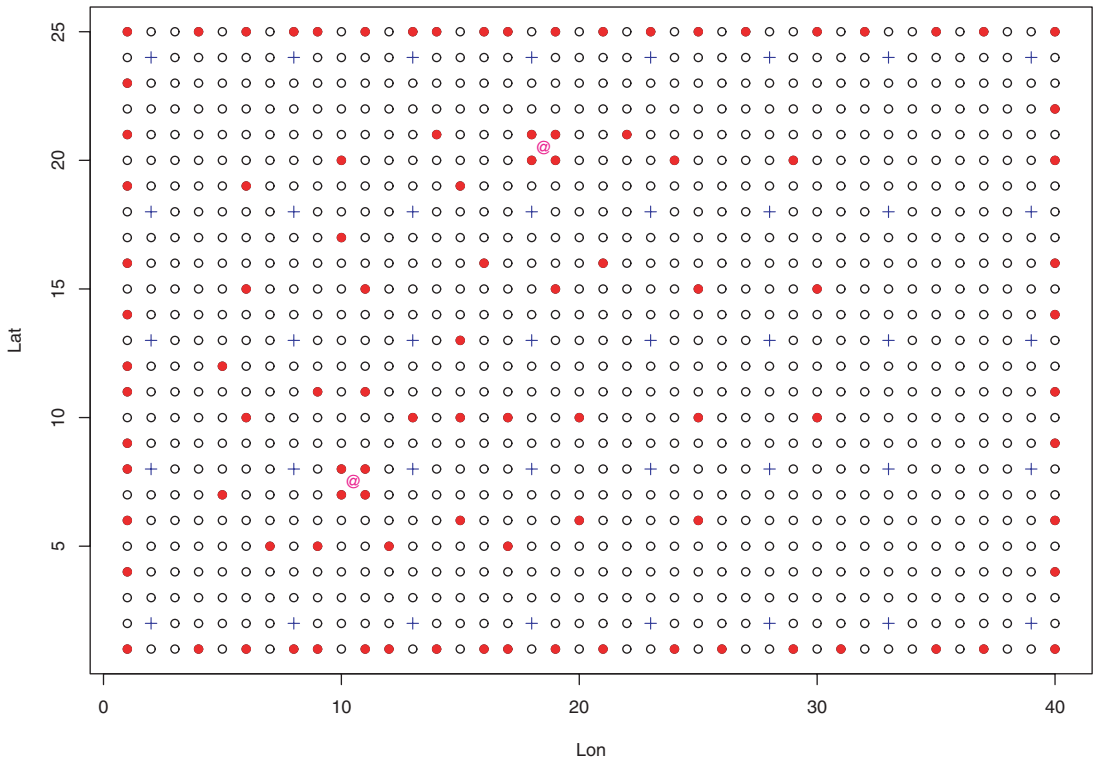

Figure 11. Sequential design based on $\boldsymbol{Y}^{*}\mathbf{I}(\beta_0, \beta_1, \beta_2)$

excessive. We have also suggested utility-weighting as a mechanism for oversampling to achieve specific objectives.

We have developed the approaches in the setting of data from a Gaussian process. However, we can work with non-Gaussian models for the data. In fact, we can also handle discrete data, for example, binary or count data by representing our model in a hierarchical fashion with the process specification moved to the second stage. (See, e.g., Diggle *et al*., 1998 or Banerjee *et al*., 2004). In these cases we merely replace the Gaussian likelihood with a different first stage likelihood before calculating the information.

A longer view of the exposure data collection might introduce a temporal component in the sense that we may seek to revisit locations that have been previously sampled at a future point in time. If we introduce suitable dependence into our modeling, we can extend our information-based sampling approaches to accommodate this setting as well.

Lastly, the foregoing development is described in terms of the spatial surface of levels for a single contaminant. A broader experiment may consider multiple contaminants. If so, we can optimize location selection when levels of several contaminants are sampled at a given location.[2] In particular,

[2]As a variant, we may have multiple types of sampling, for example, ambient sampling, ground deposition sampling, or organism sampling. Similar to the above, we can optimize sampling when multiple types of sampling will be carried out at a location.

suppose at each parcel we measure levels of say r contaminants. Now, we replace $\boldsymbol{Y}(s_i)$ with an $r \times 1$ vector $Y(s_i)$. The resulting information gain now depends on both the spatial dependence across locations as well as the dependence between the measurements within each location. A simplified form arises under a separable specification for this error structure (see, e.g., Banerjee *et al*., 2004 and references therein). The resulting form is the above information multiplied by the within location covariance matrix. Since the latter is free of n, we can use the same criteria as above in this case.

ACKNOWLEDGEMENTS

This work was supported in part by the Center for Geospatial Medicine under NIH 1 P20 RR020782-01 and under NIH 1 R21 ES013776-01. The authors thank Nils Hjort, Alicia Overstreet, and Jason Duan for helpful discussion and computation.

REFERENCES

Arbia G, Lafratta G. 2002. Anisotropic spatial sampling designs for urban pollution. *Applied Statistics* **51**: 223–234.
Assunção RM. 2003. Space varying coefficients models for small area data. *Environmetrics* **14**: 453–473.
Banerjee S, Carlin BP, Gelfand AE. 2004. *Hierarchical Modeling and Analysis for Spatial Data.* Chapman & Hall, CRC: London.
Bernardinelli L, Montomoli C. 1992. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine* **11**: 983–1007.
Bogaert P, Russo D. 1999. Optimal spatial sampling design for the estimation of the variogram based on a least-squares approach. *Water Resources Research* **35**: 1275–1289.
Brimkulov U, Krug G, Savanov V. 1986. *Design of Experiments for Random Fields.* Nauka: Moscow.
Brown PJ, Le ND, Zidek JV. 1994. Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics* **22**: 489–509.
Clayton DG, Kaldor JM. 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**: 671–681.
Cox DR, Hinkley DV. 1974. *Theoretical Statistics.* Chapman-Hall: London.
Cox DR, Reid N. 1987. Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society B* **49**: 1–39.
Diggle PJ, Tawn JA, Moyeed RA. 1998. Model-based geostatistics (with discussion). *Journal of Royal Statistical Society C* **52**: 299–350.
Dolinoy DC, Miranda ML. 2004. GIS modeling of air toxics releases from TRI-reporting and non-TRI-reporting facilities: impacts for environmental justice. *Environmental Health Perspectives* **112**(17): 1717–1724.
Fedorov VV. 1996. *Design of Spatial Experiments, Model Fitting and Prediction.* Oak Ridge National Laboratory.
Gelfand AE, Banerjee S, Gamerman D. 2005. Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* **16**: 1–15.
Guttorp P, Meiring W, Sampson P. 1994. A space-time analysis of ground-level ozone data. *Environmetrics* **5**: 241–254.
Harville DA. 1997. *Matrix Algebra for a Statistician's Perspective.* Springer: New York.
Huerta G, Sanso B, Stroud JR. 2004. A spatio-temporal model for Mexico City ozone levels. *Applied Statistics* **53**: 231–248.
Knorr-Held L. 2002. Some remarks on Gaussian Markov random field models for disease mapping. In *Highly Structured Stochastic Systems*, Green PJ, Hjort NL, Richardson S (eds). Oxford University Press: Oxford.
Le N, Zidek JV. 1992. Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *Journal of Multivariate Analysis* **43**: 351–374.
McBratney AB, Webster R, Burgess TM. 1981. The design of optimal sampling schemes for local estimation and mapping of ragionalized variables. I. Theory and method. *Computer & Geosciences* **7**(4): 331–334.
Miranda ML, Dolinoy D, Overstreet MA. 2002. Mapping for prevention: GIS models for directing childhood lead poisoning prevention programs. *Environmental Health Perspectives* **110**(9): 947–953.
Müller WG. 2001. *Collecting Data: Optimum Design of Experiments for Random Fields*. Springer: New York.
Müller WG, Zimmerman DL. 1999. Optimal designs for variogram estimation. *Environmetrics* **10**: 23–37.
Nychka D, Saltzman N. 1998. Design of air-quality monitoring networks. In *Case Studies in Environmental Statistics, Lecture Notes in Statistics*, Nychka D, Cox L, Piegorsch W (eds). Springer Verlag: New York.
Pukelsheim F. 1993. *Optimum Design of Experiments.* Wiley: New York.

Rao CR. 1973. *Linear Statistical Inference and its Applications.* Wiley: New York.
Ritter K. 1996. Asymptotic optimality of regular sequence designs. *Annals of Statistics* **24**: 2081–2096.
Sacks J, Ylvisaker D. 1996. Design for regression problems with correlated errors. *Annals of Statistics* **37**: 66–89.
Sahu S, Gelfand AE, Holland DM. 2005. Spatio-temporal modeling of fine particulant matter. *Journal of Agricultural, Biological, and Environmental Statistics.* (in press)
Schmidt AM, Gelfand AE. 2003. A Bayesian coregionalization approach to multivariate pollutant data. *Journal of Geophysical Research—Atmosphere* **108**(D24): 8783.
Su Y, Cambanis S. 1993. Sampling design for estimation of a random process. *Stochastic Process Applications* **46**: 47–89.
Shaddick G, Wakefiled J. 2002. Modelling multivariate pollutant data at multiple sites. *Applied Statistics* **51**: 351–372.
Stein ML. 1999. *Statistical Interpolation of Spatial Data: Some Theory for Kriging.* Springer: New York.
Warrick AW, Myers DE. 1987. Optimization of sampling locations for variogram calculations. *Water Resources Research* **23**: 496–500.
Xia G, Hjort NL, Gelfand AE. 2005. Information growth and consistent parameter estimation under stochastic process models with regression structure. *Technical Report*, ISDS, Duke University.
Zhu L, Carlin BP, Gelfand AE. 2003. Hierarchical regression with misaligned spatiotemporal data: relating ambient ozone and pediatric asthma visits in Atlanta. *Environmetrics* **14**: 537–557.
Zhu Z. 2002. Optimal sampling design and parameter estimation of Gaussian random fields. *PhD Thesis*. Department of Statistics, University of Chicago.
Zhu Z, Stein M. 2005. Spatial sampling design for parameter estimation of the covariance function. *Journal of statistical planning and inference*. (in press)
Zidek JV, Sun W, Le ND. 2000. Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields. *Applied Statistics and Computing* **49**: 63–79.

